一种多元台风时间序列的相似性度量方法 *

黄冬梅 1，郑　霞 1，赵丹枫 1†，王丽琳 2

(1. 上海海洋大学 信息学院，上海 201306; 2. 国家海洋局东海预报中心，上海 200129)

摘　要：台风相似性度量方法的研究对防灾减灾、辅助决策等具有重要意义。目前，台风相似性的研究大多集中在台风路径的相似性度量上。首先，梳理影响台风相似性度量的多个要素，提出了基于多元时间序列的台风数据描述方法；其次，提出了台风时间序列完整性、一致性评估与修复方法；最后，针对台风时间序列的不等长问题，设计了一种基于主成分分析和动态时间弯曲距离的相似性度量方法。通过实验验证，该方法能够实现台风相似性的有效度量。

关键词：相似性度量；多元时间序列；完整性；一致性；权值计算；动态时间弯曲

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.09.0910

Similarity measure method of multiple typhoon time series

Huang Dongmei[1], Zheng Xia[1], Zhao Danfeng[1†], Wang Lilin[2]

(1. *College of Information*, *Shanghai Ocean University*, *Shanghai* 201306, *China*; 2. *East China Sea Forecast Center of State Oceanic Administration*, *Shanghai* 200129, *China*)

Abstract: The research on similarity measure of typhoon is significant to disaster prevention and auxiliary decision-making. At present, the similarity researches on typhoon mostly focus on the path. First, this paper combed the elements of similarity measure of typhoon and using multivariate time series to describe the typhoon due to the space-time and multi-factor characteristics of it. Second, it gave one method to measure and repair integrity and consistency of typhoon time series. Finally, in the view of unequal length of typhoon time series, this paper designed a method of similarity measure based on Principal Component Analysis and weighted dynamic time warping distance. Through the experiment, this method can realize effective similarity measure of typhoon.

Key Words: similarity measure; multiple time series; integrity ; consistency ; weight calculation ; dynamic time warping

0　引言

台风作为影响我国最大的海洋灾害之一，仅 2016 年，共造成 174 人死亡、24 人失踪，直接经济损失 766.5 亿元[1]，因此，研究台风对防灾减灾、辅助决策等具有重要意义。台风相似性度量是进行路径预报、灾害预报等重要手段之一，研究其可以减少人员伤亡和经济损失。但是台风数据中普遍存在质量问题，导致分析结果存在偏差，无法提供准确的信息，造成灾害预报错误，带来不必要的财产损失和人员伤亡，因此，评估台风数据质量并对其进行修复是相似性度量的重要一部分。

台风数据质量主要体现在完整性和一致性两个方面。文献[2]综述了数据完整性和一致性的研究现状，这些方法主要围绕医学领域，几乎没有涉及海洋领域。李建中[3-6]团队在数据完整性和一致性方面的研究大多针对普通的关系数据，而台风数据是一类特殊的关系数据，具有时间属性，因此这些方法不适用于台风数据。海洋领域，文献[7-9]通过抽样方法来检验海洋数据质量是否合格，并没有提出修复方法，目前，对海洋数据质量评估与修复研究还很少，而海洋数据的质量高低直接影响海洋预报等相关研究。因而，数据的广泛应用对数据质量的保障提出了迫切需求。

国内外研究工作者对时间序列进行深入挖掘，针对不同研究重点提出各种行之有效的相似性度量方法，大致可分为两类，一类是适合等长时间序列，主要包括模式距离、余弦距离和欧式距离。模式距离[10]是表示两个序列趋势的差异程度，物理概念明确，划分合理，但表示方法粗糙，结论不够精确；基于夹角余弦距离[11]计算简单，但匹配序列长度必须相同；Agrawal 等人[12]提出的欧式距离（ED），计算简单，时间复杂度是线性的，应用较多，但只适用于等长的时间序列的相似性度量。另一类

基金项目：国家科学自然科学基金资助项目(41671431)；上海市科委地方高校能力建设项目(15590501900)；国家重点研发计划资助项目(2016YFC1401907(子课题 7)

作者简介：黄冬梅（1964-），女，河南郑州人，教授，硕士，主要研究方向为海洋数据管理、信息智能处理、辅助决策研究等；郑霞（1992-），女，硕士研究生，主要研究方向为时间序列相似性度量；赵丹枫（1982-），女（通信作者），讲师，博士，主要研究方向为海洋信息化、数据库理论、图计算等(dfzhao@shou.edu.cn)；王丽琳（1979-），女，高级工程师，硕士，主要研究方向为海洋信息技术.

是适合非等长时间序列，最具代表性的是动态时间弯曲(DTW)。DTW[13]是一种通过弯曲时间轴来更好地对时间序列形态进行匹配映射的相似性度量方法。它最早是语音识别的一种主流方法，后来 Berndt 等人[14]将它用于时间序列的相似性度量。文献[15]在 DTW 的基础上提出了最大特征优先匹配原则和自适应约束窗口法，进行心电信号的相似性度量；文献[16]提出了一种基于 DTW 聚类的水文时间序列相似性挖掘方法，提高了查找相似水文时间序列的效率；文献[17]提出了一种基于 DTW 的新型股市时间序列相似性度量方法，较好地解决股市技术分析中量价关系问题；文献[18]提出了一种关节识别的方法，可以通过一些运动，有效判断关键关节。目前，利用 DTW 进行相似台风判断还很少，文献[19]利用 Hausdorff 距离来判断台风的路径相似；文献[20]通过将台风路径概化成平面上的曲线，比较台风间数值的相似性以及形状的相似性进行台风路径的相似性判断；文献[21]利用相似离度来搜索相似路径的台风，这些文献大都只研究台风单要素相似。本文研究的多元台风时间序列，其长度不同并且各要素在相似性度量的过程中重要程度不同，因此选择加权 DTW 距离进行相似性度量研究。

本文的目标是设计一个适合具有空间性、多要素、不等长等特点的数据的相似性度量方法，从而实现台风的相似性判定，为相关部门提供准确、便捷的辅助决策方法。首先定义多元台风时间序列及其描述方法，然后给出台风时间序的质量评估与修复方法，在此基础上，考虑到台风时间序列的不等长性和相似性度量过程中各个要素重要程度不同，提出一种基于加权动态时间弯曲距离的多元台风时间序列的相似性度量方法；最后，通过实验对所提方法进行有效性分析。

1　符号和定义

a）t：观察数据的时间点，（$t=1,2,\ldots,n$）

b）j：观察数据的变量个数，（$j=1,\ldots,m$）

c）MD：移动方向

d）MV：移动速度

e）P：压强

f）Lo：经度

g）La：纬度

h）G：等级

i）V：风速

j）U：台风要素集合，$U=\{F_1,F_2,\ldots,F_m\}$

k）μ：一个抽象的度量函数

l）L：台风时间序列长度

m）Q_{AC}：完整性系数阈值

n）Q_{AU}：一致性系数阈值

o）$V_{\max}$：近中心近地面 1 min 平均最大风速（单位：节）

p）P_c：台风中心最低气压（单位：百帕）

定义 1　多元台风时间序列。一系列按时间顺序排列的台风各要素观测值 $S_t(j)$ 称为多元台风时间序列，即

$$A(U)=\begin{Bmatrix}\{S_1(1),\ S_1(2),\ldots,\ S_1(j)\},\\ \{S_2(1),\ S_2(2),\ldots,\ S_2(j)\},\\ \vdots\\ \{S_t(1),\ S_t(2),\ldots,\ S_t(j)\}\end{Bmatrix},\quad t=1,2,\ldots,n;\ j=1,2,\ldots,m$$

本文通过主成分分析法（PCA）计算以及专家意见，确定利用 MD、MV、P、Lo 和 La 5 个要素描述台风，具体表示如下：

$$A(U)=\begin{Bmatrix}\{MD_1,\ MV_1,P_1,\ La_1,Lo_1\},\\ \{MD_2,\ MV_2,P_2,\ La_2,Lo_2\},\\ \vdots\\ \{MD_n,\ MV_n,P_n,\ La_n,Lo_n\}\end{Bmatrix}$$

定义 2　要素依赖。对于台风的所有要素中任意两个要素 F_1，F_2，如果 F_1 已知，则可以计算出 F_2，则称 F_2 依赖于 F_1，记作：$F_2\to F_1$。台风等级 G 依赖风速 V，记作：$V\to G$。

2　台风时间序列质量评估及修复策略

本章针对台风数据中普遍存在的数据不完整及数据不一致现象，研究了台风数据完整性和一致性度量问题；针对台风数据完整性及一致性修复问题，设计了基于要素依赖、近邻值和其他的台风数据修复方法。

2.1　台风时间序列质量评估方法

2.1.1 台风时间序列完整性评估方法

该台风数据完整性评估方法使用如下 3 个概念：台风要素完整性、台风点完整性、台风序列完整性。其定义如下：

a）台风要素完整性。台风要素完整性是指一个要素在时刻 t 时的完整程度。对于 A 中任意元组 S 和要素 F，要素值 $S[F]$ 的完整性记作 $C_{FC}(S[F])$，可以表示为 $S[F]$ 的函数，$\mu(S[F])$，即 $C_{FC}(S[F])=\mu(S[F])$，根据不同的应用，μ 可以具有不同的形式。本文中函数 μ 可定义为

$$\mu(S[F])=\begin{cases}1,\ S[F]\text{不为空值}\\ 0,\ \text{其他}\end{cases}$$

b）台风点完整性。台风点完整性是指时刻 t 时台风所有要素的完整程度。对于 A 中任意时刻 t 的数据 S，S 的完整性记作 $C_{SC}(S)$。S 的完整性可以由 S 中的要素值的完整性来判断。则 $C_{SC}(S)$ 可以定义为

$$C_{SC}(S)=\begin{cases}0,\ S\text{中主要素}F\text{均为空}\\ 0.2,\ S\text{中主要素}F\text{有一个非空}\\ 0.4,\ S\text{中主要素}F\text{有两个非空}\\ 0.6,\ S\text{中主要素}F\text{有三个非空}\\ 0.8,\ S\text{中主要素}F\text{有四个非空}\\ 1,\ S\text{中主要素}F\text{均非空}\end{cases}$$

c）台风序列完整性。台风序列完整性是指一条台风数据的完整程度。对于任意一条台风数据 A，A 的完整性记作 $C_{AC}(A)$。A 的完整性可以由 A 中时刻 t 的数据 S 的完整性来判断。则 $C_{AC}(A)$ 可以定义为

$$C_{AC}(A)=(N_1P_1+N_2P_2+\ldots+N_nP_n)/L$$

其中：P_i 是时刻 t 的数据 S 的完整度，N_i 是完整度为 P_i 的数据 S 的个数。当 $C_{AC}(A)\geq Q_{AC}$ 时，该台风时间序列完整性可修复的。

2.1.2 台风时间序列一致性评估方法

该台风数据一致性度量框架使用如下 2 个概念：台风点一致性、台风序列一致性。其定义如下：

a）台风点一致性。台风点一致性是指时刻 t 时台风数据中不包含语义错误或相互矛盾的数据。对于 A 中任意时刻 t 的数据 S， S 的一致性记作 $U_{SU}(S)$， $U_{SU}(S)$ 可定义为

$$U_{SU}(S)=\begin{cases}0，S\text{中等级和风速相互矛盾}\\1，\text{其他}\end{cases}$$

b）台风序列一致性。台风序列一致性是指一条台风数据中不包含语义错误或相互矛盾的数据。A 的一致性可以由 A 中时刻 t 的数据 S 的一致性来判断。则 $U_{AU}(A)$ 可以定义为

$$U_{AU}(A)=N_U/L$$

其中: N_U 是满足一致性的数量，当 $U_{AU}(A)\geq Q_{AU}$ 时，该台风时间序列一致性可修复的。

2.2　台风时间序列修复方法

根据文献[22]可知，完整性错误修复结果会引起一致性、时效性、精确性的变化，因此，本文按照数据完整性修复、一致性修复的顺序对台风数据进行修复。

台风等级和风速两个要素之间存在依赖关系，本节考虑要素依赖进行完整性修复；对于无法使用要素依赖修复法的要素，本文主要考虑风速、移向、移速和压强四个要素，利用邻近值之间的关系进行完整性修复；对于无法使用邻近值修复的压强和风速，利用风压关系[23]进行完整性修复；无法使用邻近值修复的移向和移速，根据经纬度、观测值时间间隔进行修复。由于台风时间序列的一致性错误仅可能发生在等级与风速之间，本文利用要素依赖进行一致性修复。

具体算法如下：

输入：　存在完整性和一致性错误的台风时间序列 $A(U)$

输出：修复后的台风时间序列

for each　$A(U)$　do

//根据要素依赖进行等级完整性修复

if(G==null&&V!=null)

　$G=G_i$　// G_i 为根据台风风速确定的台风等级

//根据邻近值进行风速、移向、移速、压强完整性修复

if(V_i==$null$)

$$V_i=\begin{cases}\lceil (V_{i-1}+V_{i+1})/2\rceil, & i=2,3...n,V_{i+1}>V_{i-1}\\ \lfloor (V_{i-1}+V_{i+1})/2\rfloor, & i=2,3...n,V_{i+1}<V_{i-1}\\ V_2, & i=1\end{cases}$$

if(MD_i==$null$)

$$MD_i=\begin{cases}\left\lceil \frac{(MD_{i-1}+MD_{i+1})}{2}\right\rceil, i=2,...,n,MD_{i+1}>MD_{i-1}\\ \left\lfloor \frac{(MD_{i-1}+MD_{i+1})}{2}\right\rfloor, i=2,...,n,MD_{i+1}<MD_{i-1}\\ MD_2, \quad i=1\end{cases}$$

if($MV_i==null$)

$$MV_i=\begin{cases}\left\lceil \frac{MV_{i-1}+MV_{i+1}}{2}\right\rceil, i=2,...,n,MV_{i+1}>MV_{i-1}\\ \left\lfloor \frac{MV_{i-1}+MV_{i+1}}{2}\right\rfloor, i=2,...,n,MV_{i+1}<MV_{i-1}\\ MV_2, \quad i=1\end{cases}$$

if(P_i==$null$)

$$P_i=\begin{cases}\lceil (P_{i-1}+P_{i+1})/2\rceil, i=2,3...n,P_{i+1}>P_{i-1}\\ \lfloor (P_{i-1}+P_{i+1})/2\rfloor, i=2,3...n,P_{i+1}<P_{i-1}\\ P_2, \quad i=1\end{cases}$$

//根据风压关系进行压强和风速完整性修复

if($P==null,V\neq null$)

$$P=1010-(V/6.7)^{\frac{1}{0.644}}$$

if($V==null,P\neq null$)

$$V=6.7(1010-P)^{0.644}$$

//根据经纬度进行移向移速完整性修复

if($MV==null$)

$$MV=distance/temp$$

if($MD==null$)

$$\begin{aligned}\vartriangle Lo&=Lo_2-Lo_1\\ \vartriangle La&=La_2-La_1\end{aligned}$$

if($\vartriangle Lo>0\,\&\&\vartriangle La<0$)

$$angle=(90^{\circ}-angle^{'})+90^{\circ}$$

if($\vartriangle Lo\leq 0\,\&\&\vartriangle La<0$)

$$angle=angle^{'}+180^{\circ}$$

if($\vartriangle Lo\leq 0\,\&\&\vartriangle La<0$)

$$angle=(90^{\circ}-angle^{'})+270^{\circ}$$

//根据要素依赖进行一致性修复

if($V=V_i\&\&G\neq G_i$)

　$G=G_i$ //　G_i 为风速 V_i 对应的等级

end for

2.3　实例分析

本文选用“201525”号台风数据作为参考台风数据，编号为 1；再选用“201526”号台风数据和“201527”号台风数据，编号分别为 2，3，1 号台风原始数据部分如下图所示：

[{″tfbh″:″201525″,″name″:″蔷
琵″,″ename″:″Champi″,″is_current″:1,″begin_time″:″2015-10-
14T02:00:00″,″end_time″:″2015-10-25T08:00:00″,″land″:[],″points″:
[{″time″:″2015-10-14T02:00:00″,″longitude″:158.9,″latitude″:14.0,″strong″:″热
带风暴(TS)″,″power″:8,″speed″:18,″move_dir″:″西西
北″,″move_speed″:22,″pressure″:998,″radius7″:300,″radius10″:0,″radius12″:0,″ra
dius7_quad″:{″ne″:300,″se″:200,″sw″:240,″nw″:280},″radius10_quad″:
{″ne″:0,″se″:0,″sw″:0,″nw″:0},″radius12_quad″:
{″ne″:0,″se″:0,″sw″:0,″nw″:0},″remark″:″″,″forecast″:[{″sets″:″中国″,″points″:

图 1　201525”号台风部分原始数据

根据 2.2 节提出的方法，修复过程如下（以“201525”号台风为例）：

通过查询，需要进行修复的是移向和移速两个要素。具体方法如下：

a）移向修复。台风原始数据中的移动方向是采用十六风向图记录的，为了便于后续计算，需要对其进行数值表示，规定北为 0，南为 8，从北顺时针到南依次是 1—7；从北逆时针到南依次是 9—15。

根据经纬度计算出角度，对照十六风向图以及数值转换关

系补充移向。例如“2015-10-14T14:00:00”时刻，经纬度分别为156.9、15.0，后一时刻的经纬度分别为154.4、15.8，根据2.2节方法计算出角度为288.3294336171956，对照十六风向图得出移向是西西北，数值化结果为11。

b）移速修复。根据经纬度以及时间间隔计算移速。例如“2015-10-14T14:00:00”时刻，经纬度分别为156.9、15.0，后一时刻“2015-10-15T02:00:00”，其经纬度分别为154.4、15.8，计算出移速取整为28。

利用本文提出的修复方法处理之后的数据如下：

1号：$\{\{11.0,22.0,998.0,14.0,158.9\}_1,$
$\{11.0,23.0,998.0,14.3,158.2\}_2,\ldots,$
$\{11.0,23.0,990.0,15.5,152.6\}_8,\ldots,$
$\{2.0,0.0,982.0,31.1,158.0\}_{67}\}$
2号：$\{\{11.0,20.0,998.0,4.6,159.7\}_1,$
$\{11.0,20.0,998.0,5.1,158.8\}_2,\ldots,$
$\{11.0,20.0,985.0,7.7,153.4\}_9,\ldots,$
$\{2.0,40.0,998.0,21.7,137.7\}_{53}\}$
3号：$\{\{11.0,25.0,998.0,9.1,138.0\}_1,$
$\{11.0,29.0,998.0,9.4,137.3\}_2,\ldots,$
$\{11.0,19.0,990.0,11.2,132.3\}_9,\ldots,$
$\{15.0,20.0,1000.0,13.7,118.8\}_{44}\}$

图2　修复后台风数据

其中，每个元素的下标表示该点在整个时间序列中的位置，最后一个下标值表示整个时间序列的长度。

经检验，移向的均方根误差（root mean square error，RMSE）值为0.187867287325544，移速的RMSE值为0.264797369811，该修复方法精确性高。

3　基于加权DTW距离的台风时间序列的相似性度量

台风具有空间性、季节性和多要素等特点。本文综合考虑移向、移速、压强、经度和纬度五个要素，针对各要素重要程度不同且台风时间序列不等长的特点，本文采用加权DTW距离进行相似性度量。

相似性度量主要包括以下几步：判断季节相似；台风各要素权值设计；加权DTW距离计算。季节相似容易实现，不再赘述，本章节主要介绍台风要素权值计算和加权DTW距离计算两个方面。

3.1　台风要素权重设计

本文利用主成分分析法，确定描述台风的五个要素，分别是台风的移动方向、移动速度、压强、纬度和经度，它们的权重分别用W_1，W_2，W_3，W_4和W_5表示。权重计算方法是利用PCA方法计算出相关系数矩阵及其特征向量和特征值，根据设定的主成分贡献率阈值，判断主成分的个数，再根据主成分中各要素的系数计算出台风各个要素的权重。计算过程如下：

a）对原始台风数据进行标准化处理

b）计算相关系数矩阵

c）计算相关系数矩阵的特征向量和特征值

特征向量$\boldsymbol{B}$可以表示成如下所示：

$$B=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

将特征值按照从大到小的顺序排列，依次是$\lambda_1,\lambda_2,\ldots,\lambda_n$

d）选择主成分并计算其贡献率

主成分表示如下所示：

$$\begin{aligned} M_1 &= a_{11}MD + a_{21}MV + \ldots + a_{n1}Lo \\ M_2 &= a_{12}MD + a_{22}MV + \ldots + a_{n2}Lo \\ &\vdots \\ M_n &= a_{1n}MD + a_{2n}MV + \ldots + a_{nn}Lo \end{aligned}$$

每个主成分中台风要素的系数是相关系数矩阵特征向量的列，主成分的贡献率如下所示：

$$\begin{aligned} C_1 &= \lambda_1/(\lambda_1+\lambda_2+\ldots+\lambda_n) \\ C_2 &= \lambda_2/(\lambda_1+\lambda_2+\ldots+\lambda_n) \\ &\vdots \\ C_n &= \lambda_n/(\lambda_1+\lambda_2+\ldots+\lambda_n) \end{aligned}$$

e）计算台风各个要素的权值

假设主成分有n（$n\le 5$）个，则移向、移速、压强、纬度和经度的权值分别如下：

$$\begin{aligned} W_1 &= \frac{|a_{11}|}{|a_{11}|+|a_{21}|+\ldots+|a_{n1}|}\times\frac{C_1}{C_1+C_2+\ldots+C_n}+\ldots \\ &\quad+\frac{|a_{1n}|}{|a_{1n}|+|a_{2n}|+\ldots+|a_{nn}|}\times\frac{C_n}{C_1+C_2+\ldots+C_n} \\ W_2 &= \frac{|a_{21}|}{|a_{11}|+|a_{21}|+\ldots+|a_{n1}|}\times\frac{C_1}{C_1+C_2+\ldots+C_n}+\ldots \\ &\quad+\frac{|a_{2n}|}{|a_{1n}|+|a_{2n}|+\ldots+|a_{nn}|}\times\frac{C_n}{C_1+C_2+\ldots+C_n} \\ &\vdots \\ W_n &= \frac{|a_{n1}|}{|a_{11}|+|a_{21}|+\ldots+|a_{n1}|}\times\frac{C_1}{C_1+C_2+\ldots+C_n}+\ldots \\ &\quad+\frac{|a_{nn}|}{|a_{1n}|+|a_{2n}|+\ldots+|a_{nn}|}\times\frac{C_n}{C_1+C_2+\ldots+C_n} \end{aligned}$$

3.2　加权DTW距离计算

对于不同的台风，台风各要素对相似性度量的影响不同，本文利用加权DTW距离进行相似性度量，具体定义如下：

设时间序列$X=\{S_t(1),\ S_t(2),\ S_t(3),\ S_t(4),\ S_t(5)\}$，$Y=\{S_t^{'}(1),\ S_t^{'}(2),\ S_t^{'}(3),\ S_t^{'}(4),\ S_t^{'}(5)\}$，则$X$，$Y$的加权DTW距离定义见式（1）。

$$r(i,j)=\begin{cases} 0 & i=0,j=0 \\ \infty & i=0,\ j\ne 0\ or\ j=0,\ i\ne 0 \\ d(x_i,y_j)+Min & \text{else} \end{cases} \quad (1)$$

$$Min=\min\{r(i-1,j-1),r(i-1,j),r(i,j-1)\}$$

其中：对于$i=1,2,\ldots,m$，$j=1,2,\ldots,n$，用$r(i,j)$表示$r(X(1:i),Y(1:j))$，$d(x_i,y_j)$表示x_i和y_j之间的基距离，可以根据情况选择不同的距离度量，本文利用如下公式进行计算，

$$d(x_i,y_j)=\sqrt{\begin{aligned}&W_1(MD_i-MD^{'}_j)^2+W_2(MV_i-MV^{'}_j)^2\\&+W_3(P_i-P^{'}_j)^2+W_4(La_i-La^{'}_j)^2+W_5(Lo_i-Lo^{'}_j)^2\end{aligned}} \quad (2)$$

多元台风时间序列相似性度量的算法如下：

输入：历史台风数据arr2和参考台风数据arr1；

输出：历史台风与参考台风的距离d；

```
def Distance(i,j,arr1,arr2,W):
if(i==0 and j ==0):
    return 0
```

```
elseif (i==0 or j==0):
    return float("inf")
else:
    d =0
    data1 =arr1[i]
    data2 =arr2[j]
    for i in range(len(arr1[0])):
        d=d+pow(data1[i]-data2[i],2)*W[i]
    d = pow(result,0.5)
    return result+
min(Distance(i-1,j-1,arr1,arr2,W)
Distance(i-1,j,arr1,arr2,W),
Distance(i,j-1,arr1,arr2,W))
```

算法分析：设参考台风数据长度为 M ，历史台风数据长度为 N ，该算法的时间复杂度为 $O(MN)$ 。

3.3　实例分析

本节使用的数据是 2.3 节中处理之后的数据，具体计算过程如下所示（以编号 1 和 2 的数据为例）：

首先计算 1 号台风各个要素的权值，本文采用主成分分析法，得到的结果如表 1 所示。

表 1　主成分、贡献率及权值结果

	M_1	M_2	M_3	M_4	M_5	W
MD	0.1291	-0.1905	0.6530	0.1614	0.7033	0.0803
MV	0.4315	0.8247	0.3286	-0.0677	-0.1454	0.2876
P	0.8544	-0.3985	-0.1897	-0.2729	-0.0260	0.4426
La	-0.0526	0.3369	-0.5441	-0.3441	0.6851	0.0594
Lo	0.2535	0.1064	-0.3655	0.8812	0.1194	0.1301
C	0.7737	0.1824	0.027	0.0103	0.0066	

然后计算 1 号台风和 2 号台风之间的距离为

$$d\left(x_{67},y_{53}\right)=\sqrt{\begin{array}{l}W_1\left(MD_{67}-MD'_{53}\right)^2+W_2\left(MV_{67}-MV'_{53}\right)^2\\+W_3\left(P_{67}-P'_{53}\right)^2+W_4\left(La_{67}-La'_{53}\right)^2+W_5\left(Lo_{67}-Lo'_{53}\right)^2\end{array}}$$
$$=\sqrt{\begin{array}{l}0.0803\times(2.0-2.0)^2+0.2876\times(42.0-40.0)^2+\\0.4426\times(982.0-998.0)^2+0.0594\times(31.1-21.7)^2\\+0.0301\times(158.0-137.7)^2\end{array}}$$

将该结果代入式（2），得出结果为 578.4718754376307，同理，计算出 1 号台风和 3 号台风的距离为 707.0062487978438，由于 1 号台风和 2 号台风之间的距离更小，所以这两条台风更相似。

4　实验结果与分析

4.1　实验环境与实验数据

实验环境为 MATLAB R2013a，IntelliJ IDEA 2016.3.4.lnk，Win 7 SP1，1 TB 硬盘，8 GB 安装内存，Intel(R) Core^{TM} i7-3770 CPU。

选取 1945—2016 年的台风数据作为研究对象。台风数据包括台风的中英文名称、发生时间、结束时间、经纬度、强度、风速、移动方向、移动速度、压强等要素的值。

4.2　实验结果与分析

相似台风是指两条台风季节相似及满足加权 DTW 距离要求。随机选择一条台风数据为参考数据，编号为 1，再将收集到的历史台风数据分别编号为 2，3……，N （N 表示历史台风数据的数量）。如果编号 2,3… N 都满足季节相似，则计算其与参考数据之间的距离，记为 $d_{1,2}$ ，$d_{1,3}$ ……，$d_{1,N}$ ，如果距离越小，则两条台风越相似。本文设计了 3 个实验，实验一台风时间序列质量评估与修复；实验二是多元台风时间序列相似性度量方法的有效性验证；实验三是多元台风时间序列相似性度量方法的实用性证明。

实验 1　台风时间序列质量评估与修复

本实验包含两组实验，第一组实验是对需要进行实验的数据进行完整性评估及修复，第二组实验是对本文提出的台风时间序列修复方法进行有效性验证。

1）台风时间序列完整性评估与修复

（1）完整性阈值确定

实验数据采用“195526”、“200008”、“200917”、“201117”四条台风数据。利用本文提出的方法进行完整性评估，“200008”号（编号 3）台风的完整性是 0.6，“200917”（编号 4）号台风的完整性约为 0.8，将“195526”号（编号 2）台风数据进行修改，使其完整性为 0.4，分别计算三条台风与“201117”号（编号 1）台风之间的 DTW 距离，如表 2 所示。

表 2　考虑路径相似的距离（“201117”）

编号	加权 DTW 距离
1,2	37806.25047811353
1,3	45169.845710044625
1,4	77485.42226788626

由表 2 可知，与“201117”号台风相似度从大到小的排列是“195526”号、“200008”号和“200917”号，而已知“200008”号台风与“201117”号台风更相似，所以存在错误，即在数据补充过程中，完整性低于 0.6，修复的数据错误率较高。故本文中，只有当完整性大于等于 0.6 时，修复数据进行相似性度量才更准确。

（2）台风时间序列完整性评估

本实验选取的数据是下面两个实验需要使用的数据，分别是“201117”号台风、“200008”号台风、“195526”号台风、“199327”号台风、“200917”号台风、“201323”号台风、“197010”号台风、“201312”号台风、“200713”号台风、“200414”号台风、“195615”号台风、“199112”号台风、“195316”号台风、“198510”号台风和“199111”号台风 15 条台风数据。这些数据主要存在完整性问题，修复前后台风数据完整性对比图如图 3 所示。

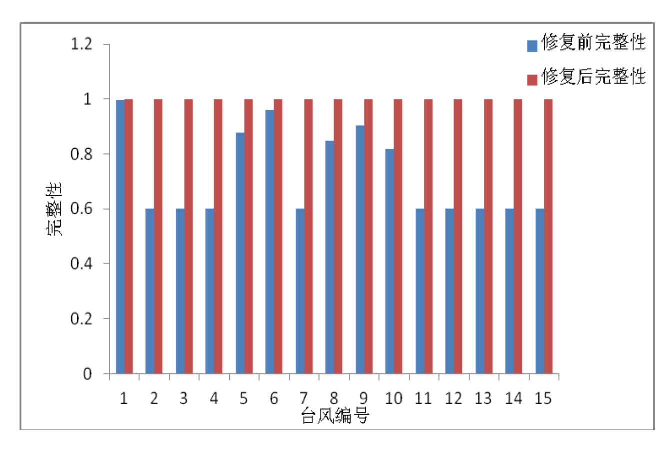


图 3　修复前后台风时间序列完整性对比

由上图可知，修复之后台风时间序列的完整性是 100%，便于后续台风相似性度量，提高相似性度量的准确率。

2)台风时间序列修复方法精确性验证

实验数据采用“200713”、“200817”、“200917”、“201011”、“201117”、“201215”、“201312”、“201323”、“201416”和“201521”10 条台风数据。

（1）完整性修复方法精确性验证

根据文献[26]可知，修复精确性可用如下公式度量：

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i-\tilde{e}_i)^2}$$

其中：e_i 是台风要素原有的真实值，$\tilde{e}_i$ 是修补的值，m 是台风的长度，$RMSE$ 的值越小，说明精确性越高，反之，精确性越低。

将原来完整的数据随机抽取 10%，用 null 值代替，利用本文提出的方法进行修复，采用邻近值修复修复方法的精确性如图 4 所示。

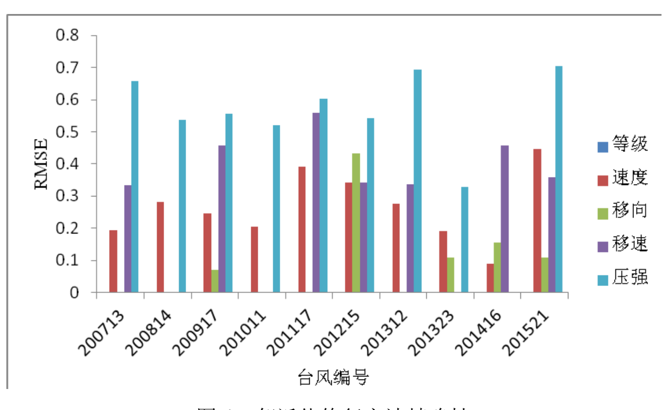


图 4　邻近值修复方法精确性

将原来完整的数据随机抽取 10%，抽取的数据每组至少有两个 null 值是相邻的，将其中的压强和风速用 null 值代替，利用风-压关系修复方法进行修复，其修复精确性如图 5 所示。

将原来完整的数据随机抽取 10%，抽取的数据每组至少有两个 null 值是相邻的，将其中的移向移速用 null 值代替，利用经纬计算修复方法进行修复，其修复精确性如图 6 所示。

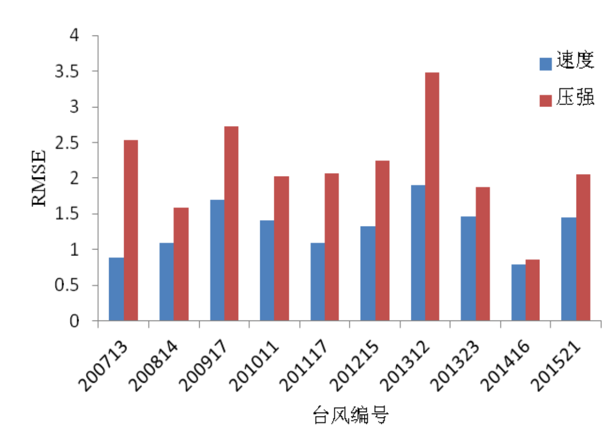


图 5　风压修复方法精确性

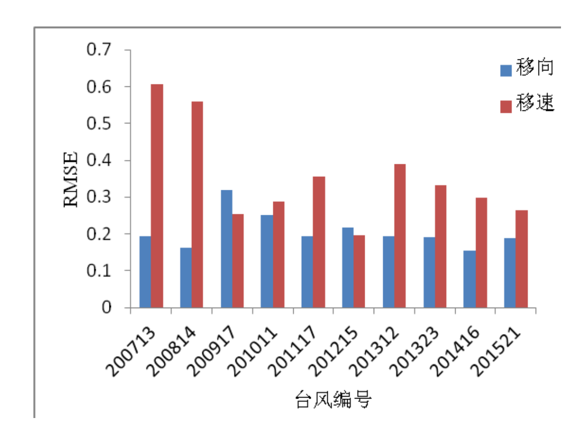


图 6　经纬计算修复方法精确性

（3）一致性修复方法精确性验证

随机抽取原来正确数据的 10%，将等级数据改成与风速不一致的数据，得到的一致性修复方法的精确性如图 4 所示。

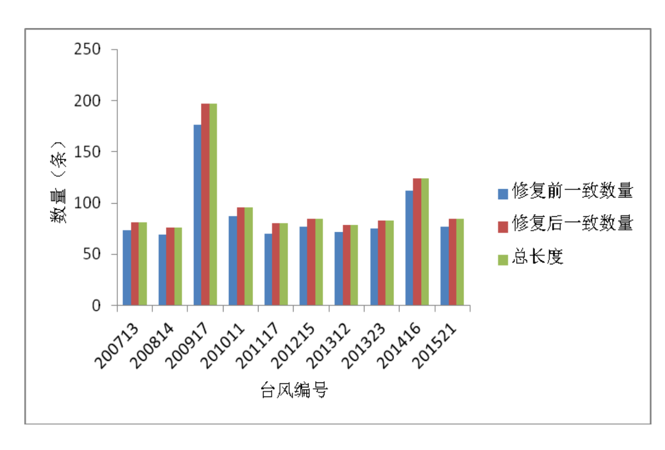


图 7　一致性修复方法的精确性

由图 4 可知，每个要素的 RMSE 的值都低于 0.8；图 5 可知，风速和压强的 RMSE 的值均小于 4，图 6 可知，移向移速的 RMSE 的值均小于 0.7，所以，本文提出的完整性修复方法修复方法精确性较高。

实验 2　多元台风时间序列相似性度量方法有效性验证

本实验包含两组实验，第一组数据是“201117”号台风数据及其相似路径台风数据，第二组数据是“201323”号台风数

据及其相似路径台风数据。利用本文的方法，只考虑路径相似，即移向移速两个要素，计算加权 DTW 距离。

1）与“201117”号台风路径相似的台风

本实验设置贡献率的阈值为 90%，权值计算结果如表 3 前两列所示。

表 3　201117”号与“201323”号主成分、贡献率及权值结果

	主成分	权值	主成分	权值
移向（*MD*）	0.0100	0.0099	0.2257	0.1881
移速（*MV*）	0.9999	0.9901	0.9742	0.8119
贡献率（*C*）	0.9967		0.9212	

选择“201117”号台风数据为参考数据，编号为 1，由文献[26]可知，与“201117”号台风相似的台风按照相似度从高到低依次是“200008”号，“195526”号，“199327”号和“200917”号，分别编号为 2，3，4，5。只考虑路径相似，由表 3 可知，移向移速的权值分别为 0.0099 和 0.9901，分别计算出台风 2 与台风 1 的距离，台风 3 与台风 1 的距离，台风 4 与台风 1 的距离，台风 5 与台风 1 的距离，计算结果如表 4 所示。

表 4　考虑路径相似的距离(“201117”)

编号	加权 DTW 距离
1,2	206.98736773989432
1,3	211.7052704239524
1,4	317.9152515180596
1,5	849.5429360222836

根据计算结果可知，距离大小按照从低到高排序为 1,2、1,3、1,4、1,5，即与“201117”号台风相似台风按照相似性从高到低分别为“200008”号，“195526”号，“199327”号和“200917”号，与文献[26]给出的结果相同。

2）与“201323”号台风相似的台风

选择“201323”号台风数据为参考数据，编号为 1，根据文献[20]，与“201323”号台风相似的台风按照相似度从高到低依次是“197010”号，“201312”号和“200713”号，分别编号为 2，3，4，只考虑路径相似性，由表 3 后两列可知，移向移速的权值分别为 0.1881 和 0.8119，加权 DTW 距离计算结果如表 5 所示。

表 5　考虑路径相似的距离

编号	加权 DTW 距离
1,2	328.82862494660293
1,3	445.0503741631772
1,4	613.0328465781108

根据计算结果可知，距离大小按照从低到高排序为 1,2、1,3、1,4，即与“201323”号台风相似台风按照相似性从高到低分别为“197010”号，“201312”号和“200713”号，与文献[20]给出的结果相同。

上述两个实验表明，本方法能够有效判断台风路径相似。以上两篇文章中的方法均没有考虑台风的季节性特点，根据本文的方法，考虑季节相似，与“201117”号相似的只有“195526”号和“200917”号，与“201323”号台风相似的只有“200713”号。

实验 3　多元台风时间序列相似性度量方法实用性验证

选择“200414”号台风数据为参考数据，编号为 1，本实验设置贡献率的阈值为 90%，利用 2.1 节权重计算方法，计算出移向移速的权值分别为 0.0315，0.9685。如果只考虑路径相似，得到的 5 个相似台风按照相似度由大到小排序分别是“195615”号、“199112”号、“195316”号、“198510”号和“199111”号，将其分别编号为 2、3、4、5、6，加权 DTW 距离如表 6 所示。

表 6　考虑路径相似的距离(“200414”)

编号	加权 DTW 距离
1,2	179.7088665502276
1,3	195.154657468981
1,4	196.21462224736027
1,5	200.0512292651246
1,6	201.13529612422815

本文除了考虑路径，考虑到台风的空间性以及台风灾害的影响，还综合考虑了压强、经纬度三个要素，能够更加全面的度量相似台风，进而有效预判台风带来的灾害程度，提前在相关地区做好防护措施。移向、移速、压强、纬度和经度的权值分别为 0.0008、0.0052、0.9926、0.0012 和 0.0002，计算的加权 DTW 距离结果如表 7 所示。

表 7 综合多要素的距离

编号	加权 DTW 距离
1,2	1880.052322218788
1,3	1259.9505176627304
1,4	1840.153387085322
1,5	1099.7791635028807
1,6	1183.6785528177586

由上表可知，与“200414”号台风相似的 5 条台风相似度由大到小排序分别是“198510”号、“199111”号、“199112”号、“195316”号和“195615”号，而与“200414”号台风路径最相似的是“195615”号台风。由下图中国台风网中各个台风的路径及强度信息可以看出，“195615”号台风强度变化与“200414”号台风最不相似，“198510”号台风强度变化与“200414”号最相似。

根据文献[27]可知，“200414”号台风云娜在浙江造成的灾害主要表现在以下三个方面：a）大风导致大量房屋倒塌，造成的死亡人数占总死亡人数的 66.5%；人员被刮倒而死亡的人数占总死亡人数的 5. 5%；部分电杆吹倒压死或因触电而死的人数占总死亡人数的 3%;b）由于台风的强风和低气压的作用，使

海水向海岸方向强力堆积，潮位猛涨，造成沿海地区堤防损坏 4 059 处 563 km，堤防决口 1 222 处 88 km;c）台风强降水使 44.4 万名群众被洪水围困，还诱发了强地质灾害。文献[28]指出，“198510”号台风造成了广西持续一周的暴雨——大暴雨的天气；文献[29]指出，“198510”号台风引发了泥石流，破坏了一批水利工程。可见，“198510”号台风带来的灾害程度与“200414”号台风相似。

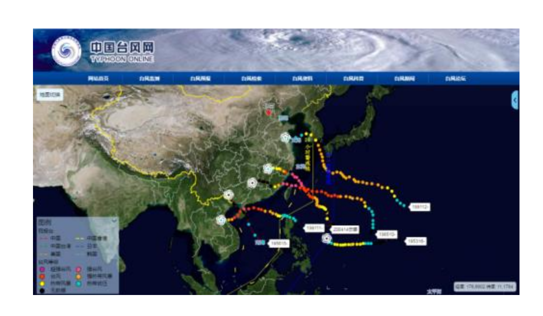


图 8　台风信息

文献[1]表明，超强台风“莫兰蒂”由于其超强强度，致使福建、 浙江、江西、上海、江苏 5 个省（市）375.5 万人受灾，44 人死亡失踪，直接经济损失 316.5 亿元。由此可见，对于台风灾害防护方面的应用，不仅需要考虑台风路径，还需要考虑其强度，因此，直接比较路径相似性是不全面的，需要综合考虑多个因素，实现台风灾害的准确预测，提前做好防护措施。

4.3　性能分析

4.3.1 时间性能对比

选择“194930”号、“195820”号、“196911”号、“197613”号、“198510”、“199215”号、“200414”号、“201117”号、“201323”号和“201521”号十条数据进行运行时间的对比，利用本文的相似性度量方法，分别考虑移向、移速 2 个要素和在此基础上，增加考虑压强、经纬度共 5 个要素所需要的相似性度量时间，得到的结果如图 9 所示。

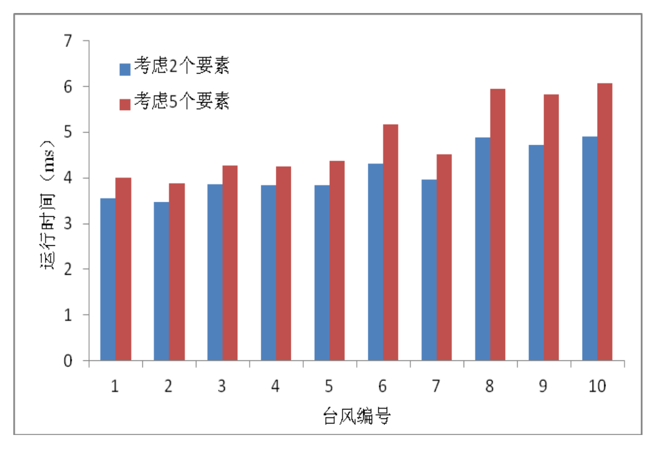


图 9　时间性能对比图

由图 9 可以看出，增加考虑压强、经度和纬度三个要素后，运行时间增加，但增加的幅度非常小。实验表明，本文的方法，考虑要素更加全面，且运行效率几乎不变。

4.3.2 参数性能对比

本文对几种相似台风的判断方法进行考虑要素、正确性以及效率方面的比较，得到的结果如下表所示，由表中结果可以看出，本文的方法在各个方面都优于其他几种方法。

表 8　参数性能比较

	路径相似性度量	考虑灾害程度	考虑季节影响	考虑地理位置	正确性	效率
基于 WebGIS 的分析系统[19]	是	否	否	是	较正确	高
相似离度原理[21]	是	否	否	否	正确	较高
基于 ArcGIS 的筛选方法[24]	是	是	否	否	较正确	高
本文方法	是	是	是	是	正确	高

5　结束语

本文首先结合台风的特点，利用多元时间序列对台风数据进行了形式化定义，实现了具有空间性、季节性以及多要素特点的数据的定量描述；对台风数据中存在的完整性和精确性问题进行评估和修复；然后以其为基础，提出了一种基于加权 DTW 距离的多元台风时间序列的相似性度量方法，实验结果表明，本文提出的台风数据质量评估方法能够提高相似性度量的正确率和灾害预测的正确性；相似性度量方法能够根据不同的参考台风设置不同的权值以适合各种相似台风的判断，能够有效判断台风造成的灾害。下一步工作是根据历史台风数据进行实时台风走势、强度等预测，为相关部门提供防灾减灾、辅助决策。

参考文献：

[1] 中国气象局. 2016 年中国气候公报. [EB/OL]. [2017-09-04]. http://m. cma. gov. cn/root7/auto13139/201705/t20170525_415102. html.

[2] 李建中，王宏志，高宏. 大数据可用性的研究进展 [J]. 软件学报, 2016, 27 (7): 1605-1625.

[3] 刘永楠，邹兆年，李建中，等. 数据完整性的评估方法 [J]. 计算机研究与发展, 2013, 50 (S1): 230-238.

[4] 苗东菁. 数据一致性的计算复杂性理论和算法研究 [D]. 哈尔滨：哈尔滨工业大学, 2016.

[5] 苗东菁，刘显敏，李建中. 概率数据库中近似函数依赖挖掘算法 [J]. 计算机研究与发展, 2015, 52 (12): 2857-2865.

[6] 刘显敏，李建中. 一种扩展条件函数依赖的发现算法 [J]. 计算机研究与发展, 2015, 52 (1): 130-140.

[7] 黄冬梅，陈括，王振华，等. 海洋数据质量检验方案中残差优化选择算法 [J]. 计算机与数字工程, 2014, 42 (10): 1752-1757.

[8] 王睿晗，黄冬梅，王振华，等. 一种针对海洋数据的空间抽样方法 [J]. 计算机应用与软件, 2015, 32 (5): 228-230+245.

[9] 王振华，周雪楠，黄冬梅. 不确定海洋数据的质量抽样检验模型研究

[J]. 计算机科学, 2015, 42 (2): 182-184+190.

[10] 朱天，白似雪. 基于模式距离度量的时间序列相似性搜索 [J]. 微计算机信息, 2007, 23 (30): 216-217.

[11] Tan P N, Steinbach M, Kumar V, et al. Wikipedia [EB/OL]. [2017-09-04]. http://zh. wikipedia. org/wiki, 2013.

[12] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases [C]// Proc of International Conference on Foundations of Data Organization and Algorithms. Berlin: Springer, 1993: 69-84.

[13] Keogh E, Pazzani M. Derivative dynamic time warping [C]// Proc of the 1st SIAM International Conference on Data Mining: RasearchGate. 2001: 1-11.

[14] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series [C]// Proc of International Conference on Knowledge Discovery and Data Mining. [S. l.] : AAAI Press, 1994: 359-370.

[15] 涂辉，刘丽，张正金. 改进 DTW 算法的心电信号相似性度量 [J]. 计算机工程与应用, 2015, 51 (16): 215-218.

[16] 杨艳林，叶枫，吕鑫，等. 一种基于 DTW 聚类的水文时间序列相似性挖掘方法 [J]. 计算机科学, 2016, 43 (2): 245-249.

[17] 冯钧，陈焕霖，唐志贤，等. 一种基于 DTW 的新型股市时间序列相似性度量方法 [J]. 数据采集与处理, 2015, 30 (1): 99-105.

[18] Reyes M, Dominguez G, Escalera S. Featureweighting in dynamic timewarping for gesture recognition in depth data [C]// Proc of IEEE International Conference on Computer Vision. 2012: 1182-1188.

[19] 孔令娜. 基于 GIS 的热带气旋路径相似法预测的研究 [D]. 长沙：中南大学, 2012.

[20] 余亮亮，何亮，缪能斌. 基于 WebGIS 的宁波市历史台风相似路径分析 [J]. 浙江水利科技, 2014, 42 (1): 24-26.

[21] 刘勇，吴必文，王东勇. 一种台风路径相似检索的算法研究 [J]. 气象, 2006, 32 (7): 18-24.

[22] 丁小欧，王宏志，张笑影等. 数据质量多种性质的关联关系研究 [J]. 软件学报, 2016, 27 (7): 1626-1644.

[23] 邹燕，赵平，乔林. 基于台风年鉴资料的台风风—压公式重建 [J]. 热带气象学报, 2009, 25 (2): 163-168.

[24] 王喜娜，黄华兵，班亚等. 利用 GIS 空间分析进行台风相似路径筛选及预测 [J]. 测绘通报, 2014 (5): 115-118.

[25] Zhang Shichao, Jin Zhi, Zhu Xiaofeng. Missing data imputation by utilizing information within incomplete instances [J]. Journal of Systems & Software, 2011, 84 (3): 452-459.

[26] 张国峰，张京红，田光辉，等. 台风灾害评估中相似台风的筛选 [J]. 湖北农业科学, 2012, 51 (7): 1334-1337.

[27] 薛根元，俞善贤，何风翩，等. 云娜台风灾害特点与浙江省台风灾害初步研究 [J]. 自然灾害学报, 2006, 15 (4): 39-47.

[28] 吴兴国. 八十年代来广西台风之特点 [J]. 广西气象, 1988, 9 (3): 6-10.

[29] 梁必骐，梁经萍. 广东台风灾害的特点及其对经济发展的影响 [J]. 中国减灾, 1993, 3 (3): 35-37, 34.